AD_____

AWARD NUMBER DAMD17-97-1-7130

TITLE: Computer-Assisted Visual Search/Decision Aids as a Training Tool for Mammography

PRINCIPAL INVESTIGATOR: Calvin Nodine, Ph.D.

CONTRACTING ORGANIZATION: University of Pennsylvania
Philadelphia, Pennsylvania 19104-3246

REPORT DATE: July 1998

TYPE OF REPORT: Annual

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1998	Annual (1 Jul 97 - 30 Jun 98)

4. TITLE AND SUBTITLE

Computer-Assisted Visual Search/Decision Aids as a Training Tool for Mammography

5. FUNDING NUMBERS

DAMD17-97-1-7130

6. AUTHOR(S)

Nodine, Calvin, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Pennsylvania
Philadelphia, Pennsylvania 19104-3246

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research And Materiel Command
ATTN: MCMR-RMI-S
504 Scott Street
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The primary goal of the project is to develop a computer-assisted visual search (CAVS) mammography training tool that will improve the perceptual and cognitive skills of trainees leading to mammographic expertise. In the first year we have carried out two experiments. The first equates experience by comparing perceptual skills of expert radiologists with laypeople searching non-medical pictorial scenes for hidden targets. Results show that expert radiology search and detection strategies do not transfer effectively to the non-medical search and detection tasks. This suggests that radiology expertise consists of specific perceptual and cognitive skills that develop primarily from experience reading medical x-ray images. In the second study, a 75-case mammogram test set was administered to 3 mammographers, 19 residents and 10 mammography techs. The results compare effectiveness of mentor-guided training on resident expertise with other levels of expertise. Not surprisingly, resident performance in detecting and classifying breast lesions was significantly inferior to experts, and no better than that of mammography techs! These findings points out the main weakness in radiology-residency training-- failure to emphasize accurate perceptual differentiation in classifying detected breast lesions. The proposed CAVS training tool is designed to remedy this shortcoming.

14. SUBJECT TERMS

Breast Cancer

15. NUMBER OF PAGES

36

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

N/A Where copyrighted material is quoted, permission has been obtained to use such material.

N/A Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

N/A Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

N/A In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

✓ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

C F Nodine 7/27/98

PI - Signature Date

TABLE OF CONTENTS (DAMD17-97-1-7130,Revised 7/31/98)

(5) INTRODUCTION:

(5.1) FOCUS. This project focuses on the training of mammography expertise which is acquired as a result of medical training and experience reading, interpreting and diagnosing breast lesions in mammograms. Mammography expertise takes years of formal training and mentoring experience with experts who read and illustrate a variety of abnormalities in breast images. Although medical training is typically rigorious and systematic, the mentoring experience during radiology residency in mammography varies widely from one teaching institution to another. Furthermore, the amount of mammography mentoring experience radiology residents receive is usually limited to four 4-6 week rotations over 4 years. This means that most radiology residents will read a total of between 1000 and 1500 mammography cases with mentor guidance during residency. We estimate that this amount of experience falls short of that of experts by a factor of from 33 to 50 times (Nodine, Kundel, Lauver, Toto, 1996). We know from testing radiology residents at the University of Pennsylvania that this amount of experience is inadequate to bring them up to diagnostic performance standards acceptable in clinical practice. For example, AFROC performance of a sample of our residents using Alternative Receiver Operating Characteristic (AFROC) methodology was A1=.653 (n= 19) compared to A1=.843 (n=3) for mammographers ($p<.01$, Sheffe test, see Resident Study).

(5.2) AIM. The primary aim of this project is to develop a mammography training tool that will, by using a computer to provide systematic feedback about visual search and detection of suspicious mammographic findings, improve the expertise of radiology residents undergoing clinical mammography residency rotation. The visual-search feedback is based on monitoring the resident's eye position during scanning of a mammogram, identifying regions containing suspicious findings based on prolonged gaze durations and highlighting these locations on the mammogram. The resident is then asked to review the highlighted areas with the help of a decision aid, determine if any abnormal features are present, and revise the original diagnostic decision. We showed in 1990 (Kundel, Nodine, Krupinski, 1990) that computer-assisted visual search (CAVS) is effective in improving the detection of lung nodules. Our goal is to apply CAVS to mammography training to see if we can enrich the experience of radiology residents during training and improve their diagnostic expertise.

(6) BODY:

(6.1) OBJECTIVES. The Statement of Work for Years 1 and 2 as stated in the proposal is as follows:

> **Technical Objective 1, Develop CAVS as a Perceptual Training Tool:** Months 1-12: Develop a computer-assisted visual search system and digital display workstation.
>
> Task 1. Program ASL Model 4000 and EYEHEAD to monitor the observer's eye position relative to head motion for digital mammography displays.
>
> Task 2. Modify eye-position data collection programs (EYEPOS/EYEDAT) to accommodate visual-dwell detection algorithm. Integrate detection algorithm with visual feedback of dwell locations on PC display workstation.

Technical Objective 2, Develop a Decision Training Tool: Months 12-24: Construct a training image set and develop a decision-aid checklist and feedback system.

Task 3. Collect mammography cases in which no lesions have been detected for at least three years and cases containing breast lesions. Digitize collected mammograms and construct the breast-lesion image training set of 160 images.

Task 4. Obtain MAMCAD and integrate training set images into MAMCAD algorithm and decision-aid checklist.

Task 5. Carry out pilot study to determine the effectiveness of the integration of the CAVS dwell-detection algorithm with the MAMCAD decision aid checklist designed to help differentiate true from false positive and negative decisions in the mammography training task.

(6.2) CHANGES IN RESEARCH PLAN. Because of a 4-month delay in recruting a research associate, and an unexpected opportunity to obtain a documented set of mammographic images from the Hospital of the University of Pennsylvania (HUP) "Technologies" database, we decided to focus Year 1's activities on Technical Objective 2 above. We have already carried out Tasks 3 and 4 under this objective. First, we have developed a 100 case test set and are in the process of verifying the status of normal, benign and malignant images (part of Task 3). We are now in the process of acquiring the HUP "Technologies" database which consists of 200 cases: 35 malignants; 65 benigns; and, 100 normals, four views each (total of 800 images). These images have been digitized using a Lumisys scanner (Model 150) at 50 microns. They will be used to supplement the original test set used in the Resident Study, and to construct a training set of 160 images to be used with the MAMCAD algorithm and decision-aid checklist (Task 4). Some of these images will also be useful for evaluating CAVS in the Objective 3.

Technical Objective 1, Tasks 1 &2, will be postponed until the training set is developed and the Resident Study written up. These two objectives will be completed by the end of grant year 2. We will submit the Resident Study to RSNA. We estimate that the HUP "Technologies" film collection, digitization and database documentation will be complete by the end of the summer of 98. During this time we will also test the computer system consisting of PCI/Dome Driver/Orwin Electronics display using Windows 95. Both the database and the PCI test are necessary prerequsites for the CAVS programming. We have already completed the preliminary testing (part of Technical Objective 1), and are now ready for CAVS programing.

The advantage of postponing the development of CAVS until year 2 will give us an opportunity to take advantage of acquiring a "clinically-proven" database of 200 mammography cases, and give the new Research Associate, Ms. Claudia Mello-Thoms, a chance to familiarlize herself with the project before programming the CAVS.

(6.3) WORK IN PROGRESS. We have submitted an article entitled "Perceptual Skill, Radiology Expertise and Visual Test Performance with NINA and WALDO" to the journal "Academic Radiology". It has been accepted for publication and will appear in the August, 1998 issue (see Nodine & Krupinski, 1998, in press, and Appendix 1).

(6.4) SUMMARY of the NINA-WALDO STUDY. This study compared visual search and perceptual analysis skills of radiologists who are expert at radiology search and detection vs. lay people using perceptual tests in which the targets of search were NINAs in

Al Hirschfeld's drawings of scenes from the theater, and WALDOs from Martin Handford's color drawings of people-cluttered scenes taken from history. We found that radiology expertise did not carry over to search and detection of these targets in the two perceptual tests. Not only did radiologists not detect more test targets than lay people in general, but they took longer than lay people on average to search and detect the NINAs and WALDOs they found as shown in Table 1.

Table 1. Mean Search Time (Sec) and Standard Deviations (SD) to First Fixate the Target in NINA and WALDO Test Pictures (n in parentheses, $p<.01$ for NINA targets).

	NINA		WALDO	
	Radiologists	Lay People	Radiologists	Lay People
Mean	16.20 (20)	9.99 (35)	26.24 (70)	22.44 (70)
SD	8.03	8.62	22.93	19.68

From these results we conclude that radiology expertise does not transfer to general search and detection tasks such as are illustrated by the perceptual tests in which NINAs and WALDOs are targets. This is true despite the similarities in perceptual-task requirements, complexity of target/background images, and signal/noise ratios. Thus, radiology expertise must require very specific perceptual and cognitive skills that develop primarily from experience reading medical x-ray images. From this experience experts most likely generate schemata of prototypic normal and various prototypic abnormals which are compared with new exemplars during image perception. Our findings suggest that these expert radiology search and detection strategies do not transfer effectively to search and detection tasks using non-medical images such as those that made up our perceptual tests.

(6.5) THE RESIDENT STUDY. As mentioned in the introduction, we are currently completing the Resident Study which provides background data on the performance of residents during mammography rotations compared to experienced mammographers and mammography techs (Nodine, Kundel, Orel, Conant, 1998, in progress). Detailed analyses of the performance data and decision time data have been performed (see below). These analyses reveal significant differences in perceptual discrimination, feature-recognition and decision-making skills among three levels of expertise which shed light on the need for systematic feedback training during the radiology residency experience. The impact of this type of training on the mammography expertise of radiology residents will be evaluated in a formal experiment after the development of CAVS.

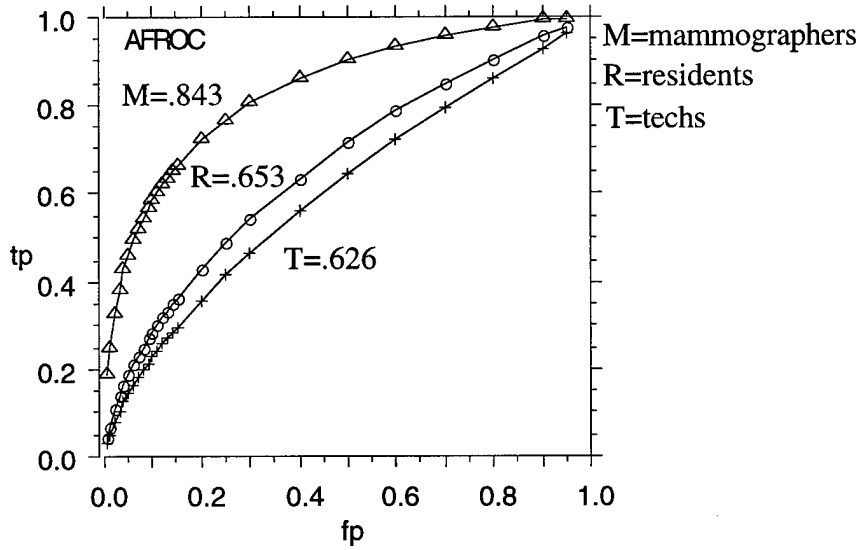
(6.6) SUMMARY OF THE RESIDENT STUDY.

PURPOSE. Mammographers need guidelines defining perceptual limits for differentiating malignant from benign lesions in order to adjust their decision making to meet the stringent clinical diagnostic demands of breast screening without sacrificing performance. We are exploring how training via clinical mammography rotation influences several aspects of perceptual performance in breast screening by comparing three levels of mammography expertise exemplified by mammographers, radiology residents and mammography techs. The research question is: How does training affect residents' accuracy in visually differentiating malignant from benign lesions in a simulated mammography screening task?

METHOD.The task consisted of 150 digital mammograms from 75 cases combined in two-view pairs. From this total of 75 image pairs there were 25 pairs containing 57 malignant lesions, 24 pairs containing 50 benign lesions and 26 pairs containing lesion-free images. This 75-pair test set was administered to 19 residents undergoing rotation and

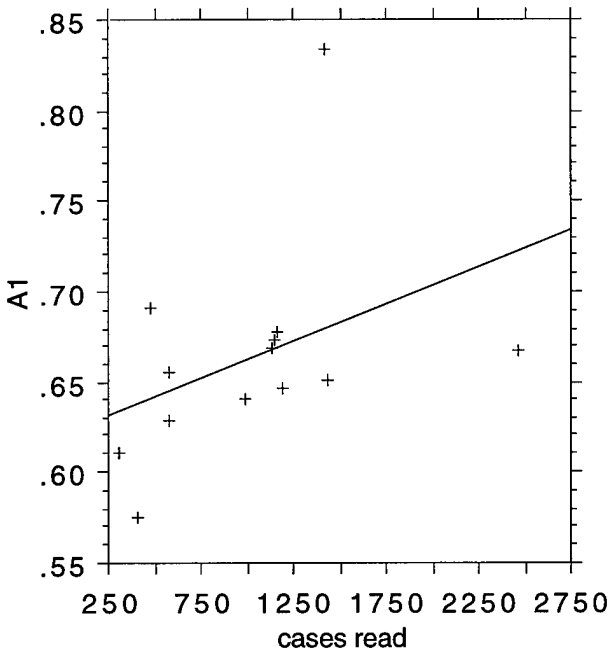
compared to 3 mammographers and 10 mammography techs. The diagnostic classification of the lesions was verified by biopsy. Normals were lesion-free for 2 years. The two-view pair of mammograms was displayed on a high-resolution (2k x2k, Tektronics) workstation. The observer interacted with the display to decide whether each image view contained (a) no malignant lesions and therefore was returned to routine screening, or, (b) suspicious lesions indicating malignancy. Depending on this initial decision, the observer either called up the next case, or, localized, classified and indicated decision confidence about each detected lesion.

RESULTS. Because multiple findings were present and lesion localization was critical for evaluation, performance was measured by AFROC and shown in Figure 1.



CONCLUSIONS. Unlike ROC, AFROC takes into account localization and classification of multiple breast findings. Our analysis points out weaknesses in training radiology expertise during resident training. The main weakness is that radiology-residency training does not stress accurate visual differentiation in localizing and classifying detected breast lesions. This is reflected in AFROC performance which, when compared to mammographers, was significantly lower for residents and equivalent to performance of techs ($p<.01$, Sheffe test). Future studies will explore the use of computer-assisted visual search (CAVS) as a training tool that provides systematic visual feedback and decision aids to improve residents' detection and classification of distinctive pathologic features that differentiate malignant from benign breast lesions.

(6.7) FOLLOW-UP OF RESIDENT STUDY. We have AFROC data on 19 residents, 3 mammographers and 10 radiology techs. By contrasting performance for these different levels of training and experience, we can gain insights into the nature of expertise. For example, we have already observed differences between the three groups in detection strategies, use of confidence ratings and lack of understanding about trade offs between TPs and FPs as a result of over-reading. Our results thus far on 19 residents seem to indicate little change in performance as a function of degree of mammography rotation mentoring experience as shown in Figure 2 (Residents only). Analysis of AFROC data shows the residents clustered around an A1= .653 (except for one resident!).



A1 = 1 + 4E-5 * cases read; R^2 = 2E-1

Performance of mammographers clustered around A1= .843. To try and understand the differences, we are beginning to look at the number of correct lesion-detection pairings on two views, CC and Oblique. The probability of detecting a pair of lesions on two views is .83 for mammographers and only .56 for residents. This suggests that residents may have difficulty predicting where to find a lesion in a second view, given a lesion detection in a first view. Eye-position data back this up (Nodine, Kundel, Lauver, Toto, 1996), but more eye-position data is needed and this is a good opportunity for Ms. Claudia Mello-Thoms to learn how to use the ASL (Applied Science Laboratories) Eye-Head monitoring system. We are currently working on transferring the localization database to the new PC computers in order to analyze differences between residents and mammographers. We also want to look at the accuracy of localizing breast lesions. We know that mammographers can localize within about 3 mm of the true target. It is important for future work with AFROC analysis to establish a localization accuracy criterion. We have yet to obtain the database and do the corresponding analysis for residents. Finally, we have an interest in developing a computer simulation using wire-mesh images of the breast to show correpondences between lesions in two views. Ms. Mello-Thoms and Mr. Toto are developing a computer-assisted teaching aid to predict lesion locations from one view to another.

Finally, in connection with the Resident Study, we have data on decision times which should prove interesting when correlated with observer confidence ratings. We hypothesize that "hard" diagnostic-image decisions take longer than "easy" decisions and preliminary data support this. We have just begun to scratch the surface in our analysis. The goal of these analyses and those indicated above is to describe differences in expertise between residents and mammographers, and develop measures of performance that can be used to quantify changes that result from CAVS training described in Objective 3.

(7) CONCLUSIONS

The primary goal of the project is to develop a mammography training tool that will improve the perceptual and cognitive skills leading to mammographic expertise. Prerequsites to this goal are an understanding of: (a) how mammographers are trained, (b) what skills are required to carry out the task of detecting, classifying and diagnosing abnormalities in mammograms, and (c) the effectiveness of current mammography training measured by evaluating the performance of residents using a test-set of mammograms representing various abnormalities. We have shifted the focus of the first-year's research from the technical development of a computer-assisted training tool (Technical Objective 1) to an analysis of the roles that training and experience play in developing image-perception skills underlying mammography expertise. To this end, we have acquired and digitized a mammogram test set of 75 cases and a mammogram training set of 200 cases, and carried out the Resident Study (Technical Objective 2). The Resident Study was designed to compare diagnostic performance across levels of expertise, and to establish a baseline performance level for residents exposed to conventional mentor-guided training which will be compared with CAVS training in project years 3 and 4.. The test set was administered to 3 mammographers, 19 residents and 10 mammography techs. The results of the Resident Study reflect on the effectiveness of mentor-guided training by comparing resident expertise with other levels of expertise. This shows, not surprisingly, that resident performance in detecting and classifying breast lesions is significantly inferior to experts. But more interestingly, it shows that resident performance is no better than that of mammography techs! This finding begins to shed light on the importance of feedback during training, and its impact on image-perception skills for detecting and classifying breast lesions. The need for systematic feedback during mammography training is suggested from these results, and the proposed CAVS provides systematic feedback which may hold the key to more effective mammography training.

(8) REFERENCES

1. Nodine CF, Krupinski EA. Perceptual skill, radiology expertise and visual test performance with NINA and WALDO. Academic Radiology 1998, in press.

2. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast lesions. Academic Radiology 1996;3:1000-1006.

3. Nodine CF, Kundel HL, Orel S, Conant E. How does training affect breast-lesion detection, localization and classification performance in digital mammogram screening? Exhibit submitted to RSNA, 1998. (Referred to as the Resident Study)

(9) APPENDICES

PERCEPTUAL SKILL, RADIOLOGY EXPERTISE AND VISUAL TEST PERFORMANCE WITH NINA AND WALDO

Calvin F. Nodine, PhD[1] and Elizabeth A. Krupinski, PhD[2]

[1]Department of Radiology, University of Pennsylvania
[2]Department of Radiology, University of Arizona

Send correspondence & reprint requests to : Calvin F. Nodine, PhD Department of Radiology University of Pennsylvania 3400 Spruce St. Philadelphia, PA 19104-6086
215-662-6630 (ph) 215-349-5115 (f) nodine@mipgsun.mipg.upenn.edu

This work was supported in part by : US Army MRMC DAMD17-97-1-7130

Running Head : Visual Test Performance

ABSTRACT

Rationale and Objectives : The goal of this study was to determine if radiologists posses superior visual search and analysis skills compared to lay people by testing radiologists and lay people on non-radiologic visual tests containing hidden targets embedded in complex pictorial scenes, but otherwise requiring no special training or experience to perform.

Materials and Methods : In two experiments, radiologists and lay people searched complex pictorial scenes (NINA and WALDO drawings) for hidden targets. Eye-position was recorded during search. Two measures of performance were obtained : accuracy of detecting targets as measured by AFROC analysis; and, visual search efficiency as measured by eye-position analysis.

Results : There were no statistically significant differences in detection performance between radiologists and lay people for either of the search tasks. Radiologists took longer on average to search the images and to first fixate the targets than did the lay people. For both groups, true-positive and false-positive decisions were associated with longer dwell times than true-negative decisions. As with radiology search tasks, false negatives were also associated with longer dwell times than the true negatives which may suggest a common perceptual differentiation mechanism.

Conclusions : Performance on two visual search and detection tasks indicate that radiologists do not possess unique or superior visual skills compared to lay people. Radiology expertise is more likely to be a combination of specific visual and cognitive skills derived from medical training and experiences searching, detecting and deciding about the diagnostic significance of findings on radiographs. Neither visual nor cognitive skills which characterize radiologic expertise carried over to NINA and WALDO tasks.

Key Words : Visual search, detection, eye-position, expertise, visual tests

INTRODUCTION

There is a common assumption that radiologists are better visual analyzers than most of their medical colleagues. Whether this visual skill is innate or acquired has been the subject of numerous studies (1 - 5). Generally, results from perceptual tests tend to correlate fairly well with general ratings of diagnostic abilities (2 - 5), but less well with specific diagnostic tasks such as pulmonary nodule detection (1). Thus, the answer to this question is unfortunately not easy to determine, primarily because innate visual skills quickly become contaminated by training and experience (6-8). Furthermore, visual testers have generally assumed that the radiologist's task is largely a visual one.

There is also a great deal of cognitive interpretation that goes into the reading of a radiograph. For example, in addition to searching for abnormalities, radiologists "read" medical images for anatomic and pathological content as they search the image. This point is largely overlooked by researchers who have developed visual tests. The radiologist's report typically contains a description the findings resulting from search, and an interpretation of the findings considered in the context of the patient's history. This separation of description from interpretation in the report provides radiologists with a framework for carrying out visual and cognitive aspects of the diagnostic radiology task in much the same way as instructions provide observers with a framework for carrying out an experimental test in the laboratory.

Artist Al Hirschfeld has been hiding the word NINA (his daughter's name) in line drawings of theatrical scenes that have appeared in "The New York Times" for over 50 years. The hide-and-seek game of finding the name NINA in Hirschfeld's drawings illustrates basic perceptual principles of detection, discrimination and decision-making commonly encountered in radiology search tasks. Hirschfeld's hiding of NINA is typically accomplished by camouflaging the letters of the name and blending them into scenic background details such as wisps of hair and folds of clothing. In a similar way, pulmonary nodules and breast lesions are camouflaged by anatomic features of the chest or breast image. Hirschfeld's hidden NINAs are sometimes missed because they are perceptually integrated into a Gestalt overview of the picture, rather than differentiated from background features during focal scanning. This may be similar to overlooking an obvious nodule behind the heart in a chest x-ray image. Because it is a search game, Hirschfeld assigns a number to each drawing to indicate how many NINAs he has hidden so as not to frustrate his viewers. In the radiologists' task, the number of targets detected in a medical image is presumed to be determined by combining perceptual input with probabilities generated from clinical history and viewing experience. Thus, in the absence of truth, searching for abnormalities in x-ray images creates opportunities for recognition and decision errors (e.g., false positives and false negatives).

Reading medical images requires both search and interpretation of radiologic findings within an anatomic image context. The task of searching, interpreting and reading the medical image uniquely combines perceptual and cognitive skills that most test developers have failed to appreciate. Interestingly, we have found experimental evidence indicating that observers have difficulty carrying out both visual search and interpretation tasks simultaneously in a testing situation. For example, in one study when observers were instructed to search for NINA, they had mixed success finding the target (9). Afterwards, the observers were asked to describe the scenes they had just searched. Interestingly, they could not describe the gist of the scene nor identify the main characters even though they were familiar well-known actors who they later identified when shown the drawings. Maybe this is why radiologists typically dictate the report while looking at the radiograph, the implication being that search has revealed findings and the image is used as a reference map during the generation of the report that both describes and interprets the findings.

The above illustration points out the need to analyze and identify task requirements before selecting tests to measure and compare radiologists' performance skills. It is clear that visual search skill is one component of the radiologist's task. Others include the ability to (a) disembed figures from

background as in hidden figures tests, (b) form an instantaneous Gestalt or global interpretation of a scene to obtain a gist and identify regions of interest for search, (c) extract distinctive features that signal perturbations in anatomic image scenery, and (d) weight the significance of distinctive features extracted from visual input during search with hypotheses generated from experience in diagnostic decision making.

We have looked for a test that taps these skills. This paper reports our experiments using two visual search tasks that come close to meeting the task requirements in radiology listed above. We compared radiologists with lay people searching art pictures to find hidden targets. The art pictures do not presuppose any prior knowledge in searching for a target. This is a way of equating observers for experience. The targets were the word NINA embedded in Al Hirschfeld's line drawings of theatrical scenes (10), and color drawings of the character WALDO embedded in people-cluttered scenic backgrounds (11). As with the anatomic scenery in radiographs, the artistically-represented scenery in our test pictures typically acts to camouflage the target, and thus the art-picture search tasks have some of the same characteristics as the radiographic search task. In addition, both medical-image targets and the art-test targets have distinctive features that provide a perceptual basis for visual differentiation of target from background. Finally, detection and recognition of targets in both medical images and test pictures are sufficiently ambiguous that observers can effectively provide confidence ratings for their decisions. Thus, we have used standard detection measures to evaluate the test results.

To summarize, this paper examines the question of what kinds of visual skills are useful to radiologists, and how training and experience influences these visual skills. We will present data from two studies comparing visual skills of radiologists with lay people on visual search tasks in which both groups are inexperienced. In each case, the subjects are required to search a picture and find a hidden target. This task is not unlike searching a chest x-ray image for a lung nodule, or mammogram for a breast lesion.

MATERIALS AND METHODS

Two types of picture search tasks were used, line drawings by artist Al Hirschfeld in which the target was the word NINA embedded in the line drawing, and color drawings by the artist Martin Handford in which the target figure was WALDO embedded among numerous colored line figures. Radiologists and lay people were recruited as observers from the University of Pennsylvania and the University of Arizona Medical Center. Five radiologists and 6 lay people from Penn served as observers for the NINA test. Seven radiologists and 7 lay people from Arizona served as observers for the WALDO test.

NINA TEST

Each observer was given a test booklet containing photocopies of 42 Hirschfeld drawings taken from "The World of Hirschfeld" book (10) containing from 0 to 7 hidden NINAs (average = 2 per picture). After an introduction and illustration of the NINA search task, observers were paced through the test booklet at the rate of 60 sec per picture to find, circle and rate confidence in detecting NINAs. A beeper sounded 10 sec before the time limit so that observers could indicate any remaining uncircled NINAs and turn the page to a new picture. When a NINA was detected and circled, observers were asked to rate their confidence in interpreting the line configuration as a NINA (3 = Definite, 2 = Probably, 1= Maybe). The number below Hirschfeld's signature that indicated how many NINAs were hidden in the picture was removed so that observers did not know how many NINAs to search for. At an average viewing distance of 40 cm, each 21.6 x 28 cm picture page subtended approximately 28 deg visual angle. The NINA targets ranged from 0.7 cm (< 1 deg) to 6 cm (> 8 deg). A chest x-ray image viewed at the same distance subtends a visual angle of approximately 42 deg and a 1 cm nodule is 1.4 deg. Eye position was monitored for a subset of 3 NINA pictures viewed by 10 observers. All of the observers had little or no experience with Hirschfeld's NINA drawings.

WALDO TEST

After an introduction and illustration of the WALDO search task, observers were shown a set of 10 full-size 48 x 31 cm color poster pictures taken from the "Where's Waldo the Magnificent Poster Book" (11). Each picture contained one WALDO plus some foils: Wilma, Wenda, Odlaw, and, numerous other characters typically reported as WALDO (i.e., false positives) ranging in size from 0.5 x 0.3 cm (1 deg) to 1.8 x 0.5 cm (3.4 deg). Observers were given up to 2 min maximum to find and point out WALDO. Feedback was given by the experimenter (EAK) as observers searched for WALDO. False positives were noted as such to the observers and the observers were told to continue searching for the real WALDO. All observers knew: who WALDO was, what he looked like, what color his clothes were, and the fact that WALDO was often partly obscured by other people/things in the picture. Confidence ratings were obtained when a WALDO was detected, however, observers did not use the entire scale and therefore the confidence rating data was discarded. The full color poster pictures subtended a visual angle of approximately 46 deg at a 30 cm viewing distance, and the WALDO targets ranged from 0.7 cm (1.3 deg) to 2.3 cm (4.4 deg).

DATA ANALYSIS

Two measures of performance were obtained: accuracy of detecting targets; and, visual search efficiency as measured by eye-position analysis. AFROC (Alternative Free Response Receiver Operating Characteristic) analysis (12) was used for the NINA task because pictures typically contained more than one target. So as not to be confused with ROC, Az, the area under the AFROC, A1, is the estimated probability of any given true target being rated higher than the most-suspicious non-target on the same image. For the NINA study, A1 was estimated from the highest rated correctly localized true positive responses on NINA relative to the highest rated false positive per picture. For the WALDO study, the observers used the rating value 6 (definitely WALDO) when they found a WALDO (a true positive) or a WALDO look-alike (false positive). They were always convinced that they had definitely found WALDO even when they were wrong! Because of this, the probability of a correct first choice localization could not be estimated. Therefore, A1 was estimated from the probability of the first correctly localized true positive response on WALDO relative to all prior false positive responses per picture.

Analysis of eye-position data focused on three measures of search efficiency: search time to fixate the target; total viewing time; and cumulative gaze duration (visual dwell). The eye position data for NINA testing was limited to a subset of 10 observers (4 radiologists and 6 lay people) and 3 pictures. Four records were lost due to poor calibration making a total of 26 records. The eye-position data from the WALDO test consisted of 7 radiologists, and 7 lay people each searching 10 pictures for a total of 140 records. The three measures of search efficiency were analyzed using t-tests and analyses of variance.

A 4000SU Eye-Tracker (Applied Science Laboratories, Bedford, MA), which records pupil and corneal reflections using an infra-red reflection source, was used to record eye position in both studies. The system has an accuracy of about 1 deg. For initial calibration purposes in the present studies, observers were seated in front the display and the observer's head was stabilized in a chin rest. After initial calibration the chin rest was removed and the observer was allowed to change position if desired. The 4000SU system comes with a head-tracker so observer head motion is recorded and integrated to adjust for eye-position changes resulting from head motion.

A detailed account of the methods used to analyze the x,y fixation data from eye-position recording can be found in Nodine et al. (13). For this study, if 50% of the area of a fixation cluster overlapped a target location (defined by an area of 0.5 deg radius surrounding the target) it was considered a "hit" (true-positive if actual target was pointed out and reported, false-negative if it was not). The same criterion was used for false-positive reports, except the fixation cluster overlapped the erroneously reported non-target location. True-negative decisions constituted those areas with fixation clusters that did not contain a target or a false positive (i.e, scenic background).

RESULTS

Table 1 shows AFROC A1 areas for finding NINA and WALDO. There was no statistical difference in the proportion of targets detected between radiologists and lay people in either task. Consistent with this finding, AFROC analysis of overall detection performance in the NINA task resulted in A1= .569 (sd = .116) for radiologists and A1= .689 (sd = .136) for lay people. The difference was not significant, t (9) = 1.58, n.s. For the WALDO task, the estimated A1 for radiologists was .650 (sd= .086) and for lay people .690 (sd= .116). This difference was also not significant, t(12)= .80, n.s.

Insert Table 1 Here

Data from eye-position recording was used to determine elapsed time until observers first fixated NINA or WALDO (true positive or false negative) after search commenced. These data are shown in Table 2. Figure 1 shows the scan pattern of a layperson, and Figure 2 shows the scan path of a radiologist both searching for NINA in a scene taken from "The Apartment" by Al Hirschfeld (10). The radiologist's search pattern contains a greater density of fixations per scanning unit (i.e., more detailed) and covers less of the image than the lay person's circumferential search pattern. This greater density was reflected in cumulative dwell for various decision outcomes which was longer for radiologists than lay people in all but one case (see Table 4).

Insert Table 2 and Figs.1 and 2 About Here

This scanning strategy difference may account for the fact that lay people were faster than radiologists fixating the NINA target, $F(1,53) = 6.93$, $p < .01$. The difference between radiologists and lay people was not significant for WALDO. This may be due to the fact that the experimenter gave observers feedback about errors during search, so that they continued to search until they either found WALDO or time ran out. Figure 3 a. shows the scene "Where's Waldo among the Monstrous Monsters" by Martin Handford (11), and Figure 3 b. shows the scan pattern of a lay person. Figure 4 a. shows the same scene. Figure 4 b. shows the scan pattern of a radiologist searching for a WALDO. The lay person repeatedly fixated WALDO (circled in the lower left corner) and reported finding it after a brief 23 sec. search. The radiologist carried out an extensive 2 min search of the scene but did not fixate or report finding WALDO.

Insert Figs. 3 a & b And 4 a & b Here

Mean total viewing time is shown in Table 3. It was shorter for lay people than radiologists in the WALDO task, but not the NINA task. Observers were given unlimited time up to 2 min to search for NINA or WALDO. There were instances in both Hirschfeld and Handford pictures where a target was not found. Because the Hirschfeld pictures contained multiple NINAs and Handford pictures contained only one, rather than try to arbitrarily adjust the viewing times by adding a constant time to reflect misses, we simply eliminated the misses from the analysis. Regardless of whether or not an arbitrary times was added into the analysis, only on WALDO pictures was mean total viewing time significantly shorter for lay people, $F(1, 110) = 5.46$, $p < .05$ (arbitrary times for misses eliminated).

Insert Table 3 Here

Table 4 shows the relationship between cumulative dwell spent on a true or false target location and the correctness of observer's decision about whether a true NINA or WALDO was or was not present at that location. Generally, observers in both NINA and WALDO tasks spent significantly longer dwelling on locations from which a positive decision (TP, FP) was generated than on locations from which a truly negative decision (TN) was generated (Sheffe test, $p < .01$). In addition, when dwelling on locations from which a falsely negative decision (FN) was generated, dwell significantly increase relative to a truly negative decision (Sheffe test, $p < .01$).

Insert Table 4 Here

DISCUSSION

There have been a number of attempts to try to correlate diagnostic ability of radiologists with a variety of perceptual tasks (1 - 5), some successful, some not. Few if any studies have compared radiologists to lay people on visual tasks that emulate what the radiologist does while searching and interpreting an x-ray image for an abnormality. We used two art-search tasks which we believed captured many of the characteristics of radiologic search, but did not require special training or experience to perform. If radiologists were either innately, or by specific training, better searchers and analyzers than lay people, the hypothesis was that the radiologists would perform better at the generalized search task. In fact, what we discovered was the radiologists were no better at the general search task than the lay people. What does this mean?

First, we assumed that the art-search tasks tap similar basic perceptual and cognitive skills of visual search, detection and interpretation as radiology tasks searching for abnormalities. This may not be the case, but before we accept this conclusion let us look at a second possibility.

Second, this study can be viewed as expanding on the nature of radiology expertise and how it transfers from one task to another. Let us assume that the art-image task may have tapped similar perceptual and cognitive skills, but that both radiologists and lay people lacked experience searching and interpreting art targets. This would have led to the same pattern of results as our first conclusion. We know from previous research that radiology expertise depends heavily on the interaction of experience with training. For example, we have shown that it takes a 13 to 200 fold increase in experience to effectively improve mammography screening performance during mammography training (6). Recent research suggests that this range of expeience may be underestimated by at least ten fold, and that during clinical mammography rotation, because of the relatively low incidence of breast cancer, radiology residents rarely encounter a case of breast cancer (14).

We know from a number of studies that radiologists search radiographs more effectively than non-radiologists. For example, re-analysis of Kundel & La Follette's 1972 study (8) shows that significantly fewer fixations were required to detect and correctly report lung lesions by radiologists and radiology residents than medical students (mean= 5.21 fixations for radiologists & residents vs. 13.27 fixations for medical students, $F(1,23)= 5.76$, $p<.05$). In this case, search efficiency is reflected by length of the scanning pattern required to sample and report the lesions correctly. This pattern of results has been repeatedly replicated (6; 8; 15; 16).

We hypothesize that because radiologists lacked experience searching for art targets, their radiology expertise did not positively transfer for the limited art-testing experience. This finding

confirms what the well-known learning theory of Osgood predicted long ago, namely, that degree of transfer depends on the similarity of training and test situations (17). The similarities in task requirements may have been outweighted by the manner in which perceptual and cognitive processes interact in finding and disembedding novel target features from art-image backgrounds compared to x-ray image backgrounds. What is critical in transfer from radiology to art tasks is the observer's understanding about how the pictorial background acts to camouflage the target, and this understanding requires a great deal of experience detecting, recognizing and deciding that a target has been found. For the radiology task of searching for a lesion in a chest or breast x-ray image, lesions are camouflaged primarily by occlusion and blending of the lesion with anatomic background structures like blood vessels on end or dense breast parenchyma. In the case of searching Hirschfeld's drawings, NINA is camouflaged by blending the letters of the name into background scenery containing features designed to mimic alphabet letters. Finally, in the case of Handford's drawings, WALDO is camouflaged primarily by mimicry. Subtle variations in the color patterns and shapes that are distinctively assigned to WALDO are also used to create foils. Thus, because different tasks call on different perceptual mechanisms for detecting and recognizing targets, what we may have observed in the present study is a low degree of perceptual-learning transfer by the radiologists so that they performed at a level of inexperienced lay people. In fact, our data show that radiologists tended to find fewer art targets and miscalled more art targets falsely than laypeople. From the standpoint of transfer of radiology expertise, neither perceptual discrimination nor visual search skills carried over to the art tasks.

Finally, analysis of eye-position data revealed that when both radiologists and lay people missed art targets, they typically spent significantly more visual dwell fixating the true target than negative, non-target background locations on the images. This finding together with the ranking of dwell times associated with true and false positive decisions has also been observed in visual search tasks in radiology (7; 18;19). Thus it seems that at least in this respect, the art-image task was tapping fundamental perceptual processes associated with visual search, detection and decision making.

These data have a couple of important implications for testing and training. The first implication follows from our conclusion about transfer: radiologists may not be superior visual searchers and analyzers in a general sense. They may be quite expert at searching radiologic images (8), but their search and analysis skills do not transfer to new tasks having similar requirements. If this is true, then this finding has direct consequences on two other situations : 1) selecting residents for radiologic training (and developing tests for this selection process); and 2) methods of training during radiology residencies. A recent study by Freundlich and Murphy (20) found that 93.5% of medical students taking a radiology elective expected that they would be able to correlate their interpretations of radiographs and other medical images with radiographic reports. But did they consider what happens when a disagreement occurs? Obviously not. Even more surprising was the finding that many medical students felt that a four-week elective adequately prepared them to independently interpret radiographs. In all probability radiology residency programs do not share this view. In fact, there are efforts being made to change the radiology residency curriculum (21 - 23) to better prepare residents for a career in radiology. The main question, of course, is exactly what and how do we teach residents to be expert radiologists. Our results suggest that perceptual skills for radiology require knowing what distinctive features differentiate abnormal from normal anatomic structures (via medical training) and how these features are transformed by x-ray imaging and interpreted within the context of diagnostic hypothesis testing and problem solving (via radiologic experience). As our results suggest, these skills may in fact be specific to the situation of interpreting radiographs, and may not generalize to other non-radiologic hide-and-seek search tasks.

The testing and training issue is also interesting in light of the fact that many training institutions may have to decrease the number of radiology residents in the near future (24). Our study raises questions about the effectiveness of testing programs to predict which medical students would make good radiologists. Our findings show how difficult it is to develop a testing situation that

predicts how much perceptual learning carries over from radiology search to visual tests. On a generalized search and analysis task, radiologists are no better than lay people. Bass and Chiles (1) found that performance on perceptual tests had little correlation with diagnostic accuracy in detecting pulmonary nodules. Berbaum et al. (2 - 5) did find a good correlation between perceptual test performance and ratings of residents' diagnostic skills, but one of the studies (3) found that the correlation was poor during the first year, and stronger only after that.

These studies differ from the present study in that they did not compare radiologists with lay people. They looked only at radiologists and those already training to be radiologists. Therefore, the effects of training may have already influenced their skills to some degree. It is impossible to tell whether the observers tested had different or better perceptual skills coming into their residency, or whether the training enhanced or fostered already existing perceptual skills that had not previously been tapped. Our study tested radiologists and lay people on a visual search task and found little difference in performance, suggesting that if radiologists do possess superior search skills they may only be specific to the radiologic search task and may not be evident on other types of search tasks that do not deal with radiologic images.

REFERENCES

1. Bass JC, Chiles C. Visual skill : correlation with detection of solitary pulmonary nodules. Invest Radiol 1990;25:994-998.

2. Smoker WRK, Berbaum KS, Luebke NH, Jacoby CG. Spatial perception testing in diagnostic radiology. AJR 1984;143:1105-1109.

3. Berbaum KS, Smoker WRK, Smith WL. Measurement and prediction of diagnostic performance during radiology training. AJR 1985;145:1305-1311.

4. Berbaum KS, Platz C. Perception testing in surgical pathology. Human Pathology 1988;19:1127-1131.

5. Smith WL, Berbaum KS. Improving resident selection : discrimination by perceptual abilities. Invest Radiol 1991;26:910-912.

6. Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Acad Radiol 1996;3:1000-1006.

7. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. Acad Radiol 1996;3:137-144.

8. Kundel HL, LaFollette PS. Visual search patterns and experience with radiological images. Radiol 1972;103:523-528.

9. Nodine CF, Carmody DP, Kundel HL. Searching for NINA. In Senders JW, Fisher DF, Monty RA (Eds.) Eye movements and the higher psychological functions. Hillsdale, NJ:LEA 1978:241-258.

10. Hirschfeld A. The world of Hirschfeld. New York, NY : Abrams, 1970.

11. Handford M. Where's Waldo? The magnificent poster book. New York, NY : Little, Brown & Co., 1991.

12. Chakraborty D, Winter LHL. Free-response methodology : Alternative analysis and a new observer-performance experiment. Radiol 1990;174:873-881.

13. Nodine CF, Kundel HL, Toto LC, Krupinski EA. Recording and analyzing eye-position data using a microcomputer workstation. Behav Res Meth, Instrum & Comput 1992;24:475-485.

14. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med 1996; 156:209-213.

15. Gale AG, Walker GE, Roebuck EJ, Wilson ARM, Worthington BS. Mammographic screening: What do radiologists look for? British Journal of Radiology 1990;63:S52.

16. C. Lesgold A, Rubinson H, Feltovitch P, Glasser R, Klopfer, D, Wang Y. Expertise in a complex skill: Diagnosing x-ray pictures. In MTH Chi, R Glaser, MJ Farr (Eds.) The nature of expertise. Hillsdale, NJ, LEA; 1988:311-342.

17. Osgood CE. Method and theory in experimental psychology. NY: Oxford University Press, 1956, 532.

18. Kundel HL, Nodine CF, Carmody DP. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Invest Radiol 1978;13:175-181.

19. Kundel HL, Nodine CF, Krupinski EA. Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. Invest Radiol 1990;25:890-896.

20. Freundlich IM, Murphy WA. Medical students who choose a radiology elective : Career decisions, motivations, and intentions. Acad Radiol 1995;2:527-532.

21. Mundy WM, Binet E. A comprehensive objective-based curriculum for radiology residents. Acad Radiol 1995;2:173-178.

22. Ekelund L, Lanphear J. Diagnostic radiology in an integrated curriculum : Experience from the United Arab Emirates. Acad Radiol 1997;4:653-656.

23. Rao VM. A perspective on radiology residency curriculum guidelines : Results of the 1995 survey of program directors. Acad Radiol 1996;3:512-516.

24. Manaster BJ. Decreasing the number of radiology residents : The impact on radiology departments and resident education. Acad Radiol 1995;2;1113-1114.

Table 1. AFROC A1 Area Values in NINA and Estimated A1 Area Values for the WALDO Test Pictures.

	NINA		WALDO	
	Radiologists	Lay People	Radiologists	Lay People
O1	.526	.566	.600	.900
O2	.511	.728	.683	.750
O3	.482	.528	.650	.650
O4	.552	.874	.683	.550
O5	.772	.639	.650	.600
O6		.802	.500	.733
O7			.783	.650
Mean	.569	.689	.650	.690
SD	.136	.116	.086	.116

Table 2 . Mean Search Time (Sec) and Standard Deviations (SD) to First Fixate the Target in NINA and WALDO Test Pictures (n in parentheses).

	NINA		WALDO	
	Radiologists	Lay People	Radiologists	Lay People
Mean	16.20 (20)	9.99 (35)	26.24 (70)	22.44 (70)
SD	8.03	8.62	22.93	19.68

Table 3. Mean Total Viewing Time (Sec) and Standard Deviations (SD) to Search for NINA or WALDO Targets in Test Pictures (n in parentheses).

	NINA		WALDO	
	Radiologists	Lay People	Radiologists	Lay People
Mean	44.90 (11)	44.66 (15)	61.42 (55)	48.02 (57)
SD	21.48	17.11	32.13	28.48

Table 4. Mean Cumulative Dwell (ms) and Standard Deviations (SD)Associated with Various Decision Outcomes for NINA and WALDO Test Pictures (n in parentheses).

		NINA		WALDO	
		Radiologists	Lay People	Radiologists	Lay People
True Positive	Mean	2525 (14)	1393 (17)	1775 (55)	1225 (57)
	SD	1315	981	1354	676
False Negative	Mean	1340 (8)	1223 (7)	2773 (15)	2046 (13)
	SD	911	825	1425	1214
False Positive	Mean	---	---	1585 (26)	1475 (20)
	SD	---	---	800	749
True Negative	Mean	798 (64)	521 (57)	937 (9736)	993 (10421)
	SD	806	599	1641	1475

FIGURE CAPTIONS

Figure 1. Scanpath of lay person searching for NINA in "The Apartment" by Al Hirschfeld. The lay person carried out a clockwise circumferential scan and fixated the NINA at 9 sec.

Figure 2. Scanpath of a radiologist searching the same scene. The radiologist's scanpath got tangled in the spaghetti being strained by Jack Lemon's tennis racket for 11.5 sec before moving on. As a result, he did not fixate the NINA until 20 sec into the search. Notice that even though the size of the NINA target is relatively large, because the letters are integrated into the structure of the lamp, the target lacks peripheral conspicuity and therefore requires direct fixation in order to be detected.

Figure 3 a. An example of a "Where's Waldo?" scene (Where's Waldo among the Monstrous Monsters?"). The drawings used in the present experiment were the full-color 48 x 31 cm poster size pictures. The reduced black-and-white photographs give a false impression of the actual search task. However, the photographs do convey the density of pictorial detail present in the original. **Figure 3 b.** The scanpath of a lay person. The lay person started search near the lower middle of the picture (designated by the triangle) and reported finding WALDO after 23 sec of search. WALDO is circled in the lower right corner of the picture and scanpath. Note the density of fixations required to search the dense pictorial detail for WALDO.

Figure 4 a. The same scene. **Figure 4 b.** The scanpath of a radiologist. WALDO (circled) is in the lower right corner of the scene. The radiologist began the search in approximately the same location as the lay person but did not find WALDO (false-negative) during the 2 min search period, even though he did fixate WALDO (as indicated by the circle) toward the end of search.

Figure 1

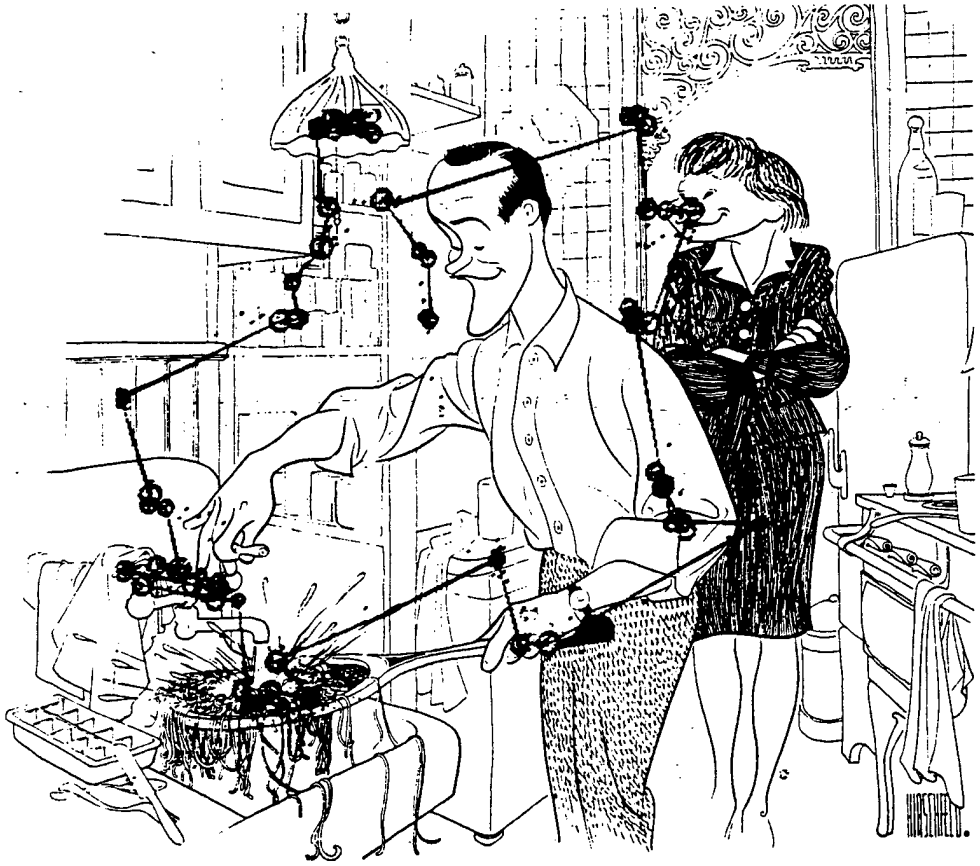
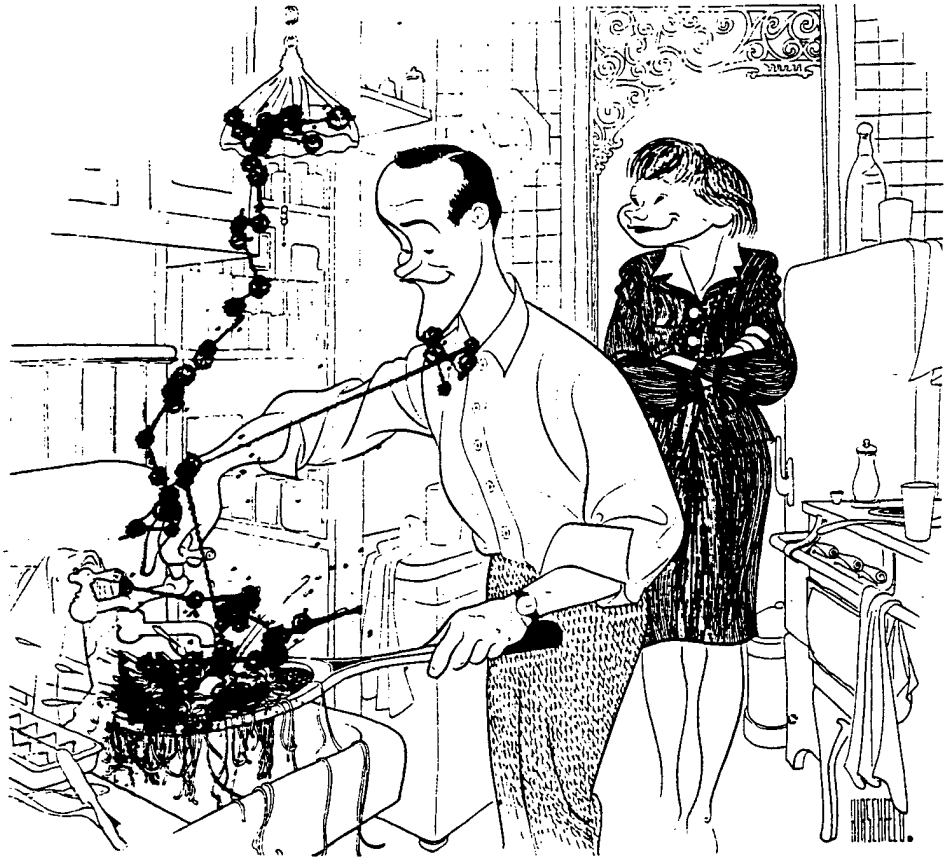


Figure 2



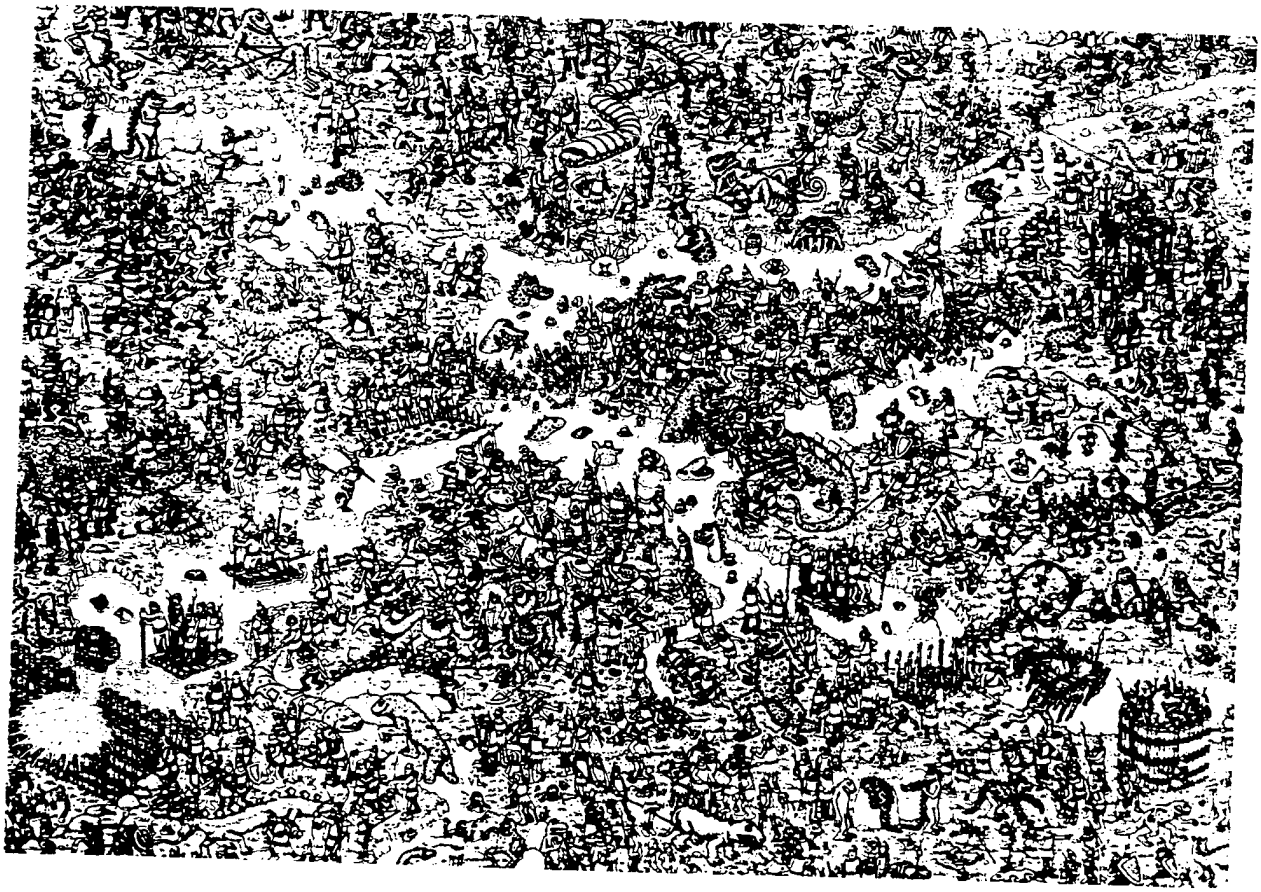


Figure 3a

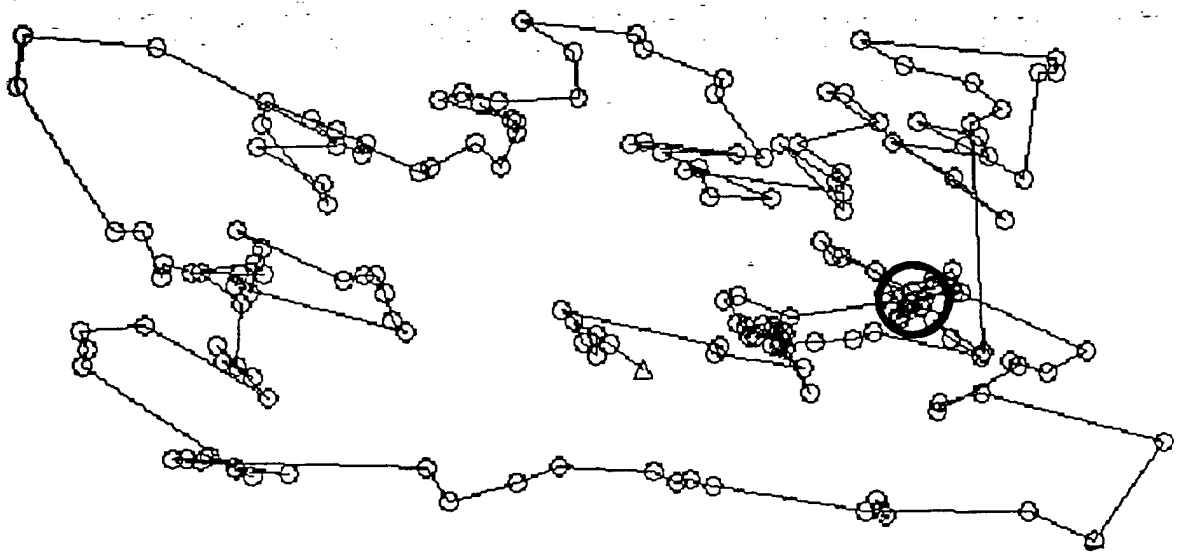


Figure 3b

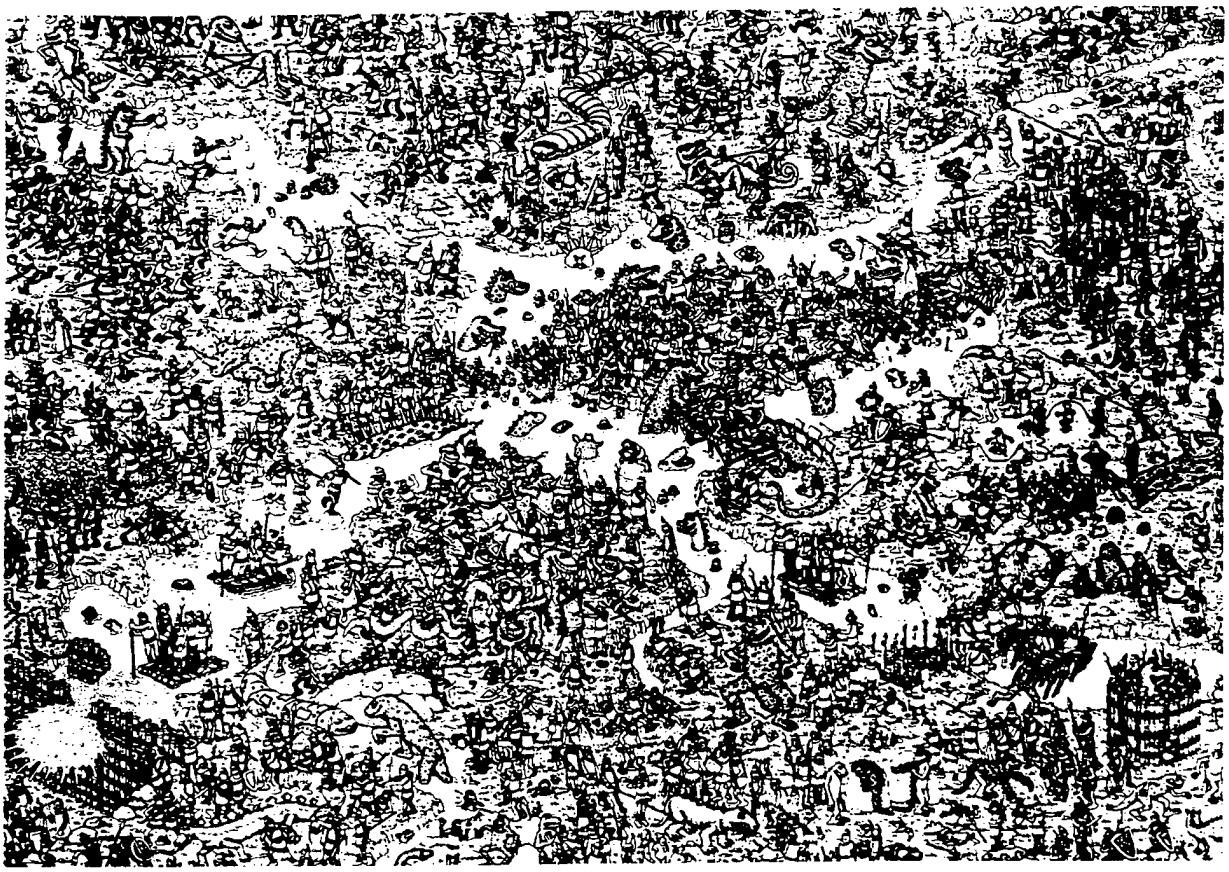


Figure 4a

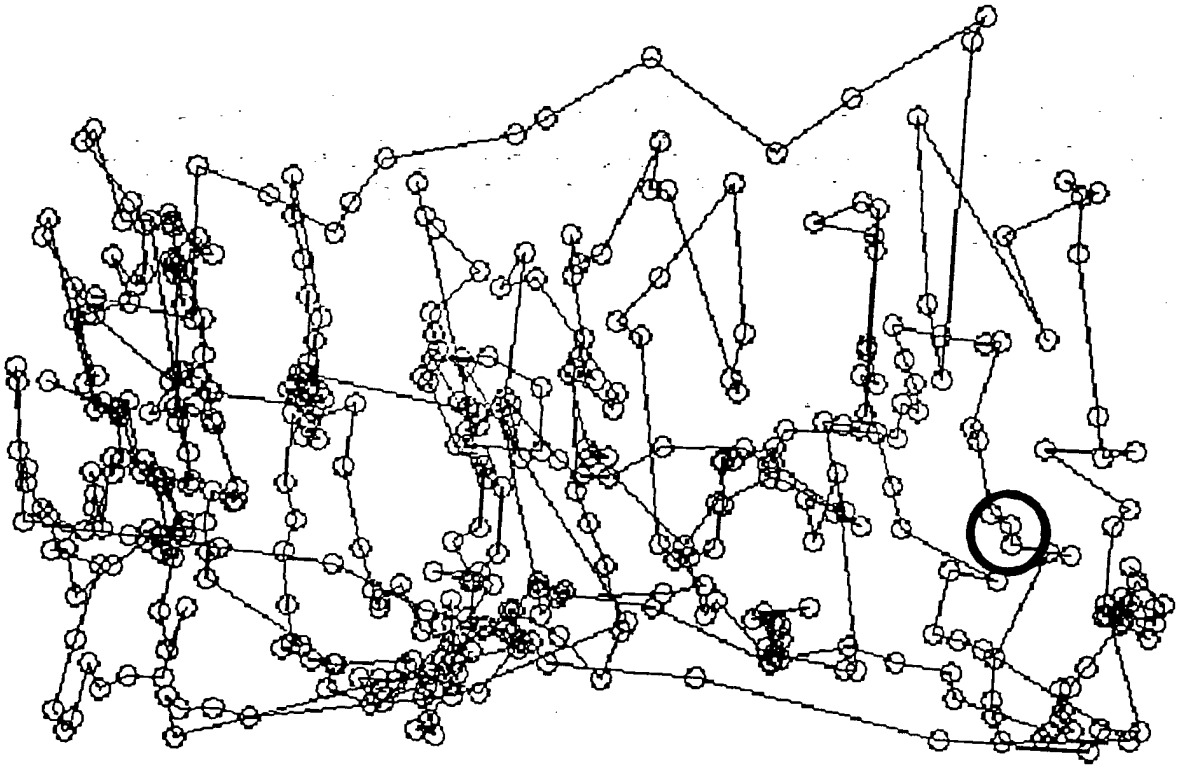


Figure 4b

PROGRESS REPORT DAMD17-97-1-7130 APPENDIX 2

Enhancing recognition of lesions in radiographic images using perceptual feedback

Elizabeth A. Krupinski, MEMBER SPIE
University of Arizona
Department of Radiology
Tucson, Arizona 85724
E-mail: krupinski@radiology.arizona.edu

Calvin F. Nodine, MEMBER SPIE
Harold L. Kundel, MEMBER SPIE
University of Pennsylvania
Department of Radiology
308 Stemmler Hall
Philadelphia, Pennsylvania 19104

Abstract. When radiologists search a medical x-ray image for an abnormality, their eyes often fixate and refixate the true target, dwelling on it for prolonged times, often without recognizing that they have discovered the object of search. Monitoring the eye position of the radiologist provides the x and y coordinates of the dwelling location. This location can be superimposed on the image and dynamically fed back to the radiologist for reevaluation. When this is done, the probability of recognizing and reporting an abnormality is shown to be enhanced significantly. An increase of 20% in observer performance is observed for radiologists searching chest images for tumors after receiving perceptual feedback compared to a second look without perceptual feedback. The true-positive rate increased and the false-positive rate decreased. Perceptual feedback represents a potentially significant technique for enhancing lesion recognition in radiology. [S0091-3286(98)00203-7]

Subject terms: observer performance; visual dwell; perceptual feedback; lesion recognition.

Paper ART-105 received May 27, 1997; revised manuscript received July 30, 1997; accepted for publication Aug. 10, 1997.

1 Introduction

In medical x-ray imaging, tumors and fractures make up a significant portion of the types of abnormalities to be detected and recognized during search. In radiology, search typically means visually scanning an x-ray image and deciding whether or not an abnormality is present. The task is a difficult one, and there is estimated to be about a 30% miss rate[1–3] in radiology, with certain types abnormalities being missed more than others (e.g., fractures in bone images). Our goal is to try and understand why radiographic abnormalities are missed and how errors can be reduced and performance increased. Eye-position analysis has been a useful tool in helping us to (1) understand how visual search is performed in radiology, (2) isolate the perceptual and cognitive causes of error, and (3) design a perceptual feedback system to enhance the recognition of missed abnormalities. One major assumption behind the use of eye-position recording is that the amount of time the eyes spend looking at an object reflects information processing, object encoding and recognition.[4–6]

2 Eye-Position Recording and Analysis

Complete details of the eye-position recording and analysis methods can be found in Nodine et al.[7] For the studies discussed in this paper, eye position was recorded either with an Eye-Trac Model 210 or an Eye-Tracker 4000SU (both from Applied Science Laboratories, Bedford, Massachusetts). Both systems operate on the same basic principle. The main difference between the two systems is that the 4000SU is capable of monitoring head movements, eliminating the need for observers to maintain their head position rigidly during eye-position recording. Both recording systems are IR based and compute line of gaze and dwell time on the basis of pupil and corneal reflection parameters.

Eye position is sampled every 1/60 s and the recording system assigns an x,y coordinate to each sample. With the 4000SU system, the head tracker and eye tracker data are integrated in real time so that the x,y coordinate data reflect this integration. Fixations are then formed by grouping the x and y coordinates using a running mean distance calculation having a 0.5 deg radius threshold. Clusters can then be formed by grouping fixations, and cumulative clusters can be formed by combining individual clusters. Typical scan patterns on normal and abnormal chest images are shown in Figure 1. The studies discussed in this paper used cumulative clusters of dwell times.[8]

3 Visual Dwell and Diagnostic Decisions

The recognition of abnormalities in radiology can be very difficult. Since the abnormality is typically hidden, it cannot be detected either by peripheral pickup or by a chance landing on the abnormality.[9] Focal scanning by the high-resolution central vision must systematically cover regions of suspicion in the image that are likely to contain abnormalities. Once a target candidate is detected, it must be visually scrutinized to integrate the imaged features into a recognizable representation of the sought-after abnormality. The process of detecting, integrating and testing a target candidate for distinctive features, and deciding whether or not to report it as an abnormality requires prolonged dwelling on the region of interest during the course of scanning the image.[10]

Figure 2 shows a model of how visual dwell is related to search, recognition and decision making. Dwell time comes

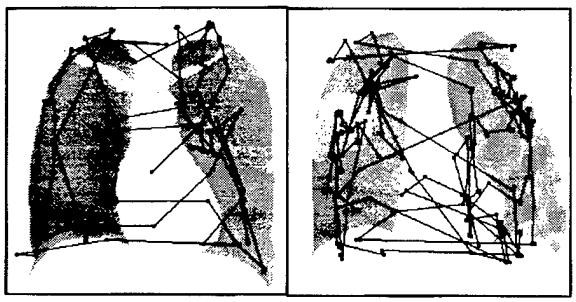


Fig. 1 Scanpaths of radiologists looking at a normal chest image on the left and an abnormal image on the right. The abnormal image contains two patches of pneumonia indicated by the light gray shading in the darker gray of the lungs. The fixations, which typically last from 200 to 300 s, are indicated by dots that are connected by lines. The lines indicate the order in which fixations are generated. The abnormal image shows a number of clusters of fixations while the normal image is free from distinct clusters.

into the picture during the analysis and testing of a target candidate for distinctive features, and can be thought of as reflecting visual information processing. Using a signal detection framework, visual dwell can be correlated with positive and negative decisions made by the radiologist.

This correlation is accomplished by relating the dwell times of cumulative cluster x and y locations with the x and y image locations of abnormalities and the reports of the viewer. If a cumulative cluster falls within 2.5 deg of a reported or missed abnormality, it is associated with a true positive (TP) or false negative (FN) decision, respectively. If a cluster falls on an abnormality-free area erroneously reported as containing an abnormality, it is associated with a false positive (FP) decision. Any cluster falling on an unreported abnormality-free location is defined by default as a cluster associated with a true negative (TN) decision.

Within this framework, we have been able to show in four separate experiments that prolonged visual dwell predicts the location of real (TP) and false (FP) abnormalities. More importantly, prolonged dwell has been shown to predict the location of missed abnormalities (FNs).

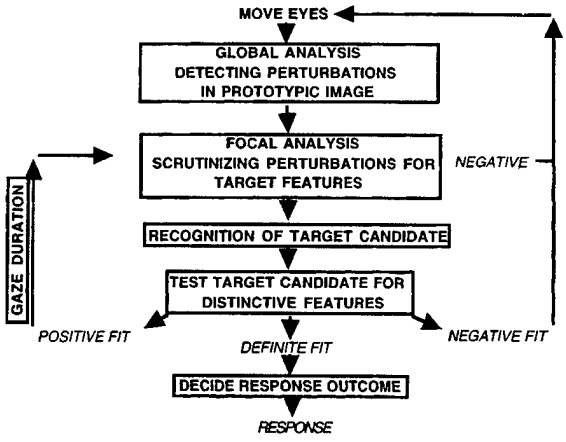


Fig. 2 Model of the relationship between scanning, dwelling and decision making for the radiologic task of searching for lesions in x-ray images.

Table 1 Median cumulative cluster dwell times (in milliseconds) associated with TP, FN, FP, and TN decisions for the chest, mammography, and bone studies.

Study	TP	FN	FP	TN
Chest	2291	1283	2091	547
Mammography	2249	1638	2003	892
Bone trauma	1286	938	895	532
Bone fractures	734	766	495	460

4 Eye-Position Recording Studies

The first of the four studies[11] used chest images with tumors as targets of search. Twelve radiologists searched 40 images (half with a single subtle tumor, half without) for 15 s each, while eye position was recorded. The second study[12] examined the eye-position data of six radiologists searching 20 mammography cases (40 images, right and left breast images of the craniocaudal or mediolateral oblique views). Fifteen of the cases had one or more masses and/or microcalcification clusters and five cases were abnormality free. Observers had unlimited search time. In the third study,[13] three bone radiologists and three orthopedic surgeons searched 27 bone images for fractures and other signs of trauma. Eighteen of the cases had subtle signs of trauma (fracture, swelling, dislocation, joint effusions, ligamentous injury) and 9 were normal. Observers had unlimited viewing time. The fourth study[14] used nine bone images, seven with subtle fractures. Fifteen observers had 30 s to search the images.

The median dwell times of cumulative clusters for each decision category for these four experiments are presented in Table 1 and the decision performance data are presented in Table 2. The survival curves[15] for the four sets of data are shown in Figure 3. Survival analysis is used to characterize the distributions of dwell times. It creates a plot that represents the percentage of cumulative clusters that fall within a given dwell interval.

The important thing to note from these data is the similarity in dwell times and the shapes of the survival curves associated with each of the decisions across the four types of images and abnormalities studied. In general, TP and FP decisions tend to be associated with the longest dwell times—taking up to 1744 ms longer than TN decisions in the case of chest images with tumors. Presumably this increased time reflects feature extraction and decision making (see Figure 1). TN decisions tend to be associated with the shortest dwell times. FNs tend to fall between these two

Table 2 TP, FN, FP, TN percentages for the four eye-position recording experiments performed.

Study	TP	FN	FP	TN
Chest	35	65	8	92
Mammography	81	19	33	67
Bone trauma	80	20	15	85
Bone fracture	70	30	41	59

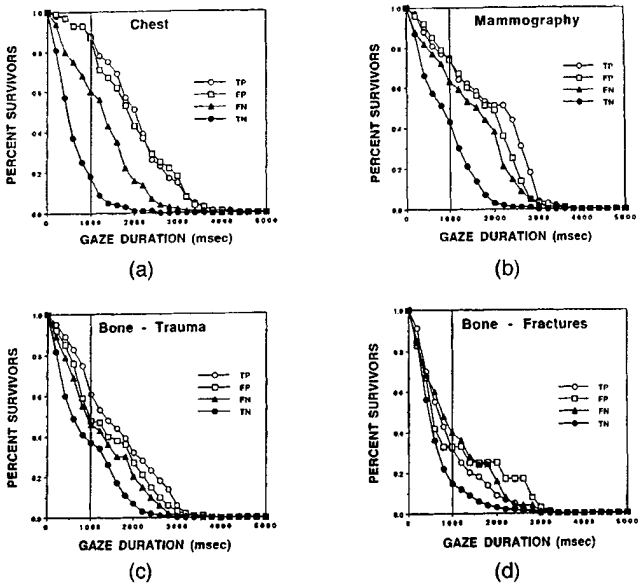


Fig. 3 Survival function curves associated with TP, FP, FN, and TN decisions for the pooled cumulative cluster gaze durations in milliseconds for the (a) chest, (b) mammography, (c) bone trauma, and (d) bone fracture eye-position recording studies. The survival function indicates the probability of survival of cumulative fixation clusters as a function of gaze duration (200 ms intervals). The vertical lines indicate the percentage of decision clusters remaining after the 1000 ms threshold.

extremes and have dwell times significantly longer than those for TNs (range=306 ms for bone fractures to 746 ms for mammography).

5 Omission Errors and Visual Dwell

Dwell data, such as that already presented above, has also been used to further classify FN errors (misses) into three main types.[16] Approximately 30% occur because of incomplete scanning of the image by the highly sensitive foveal region of the retina. These are called scanning errors, due to failure of perceptual input. Another 30% occur because of a failure in the basic mechanism of object recognition. The missed abnormality is fixated centrally and receives visual dwell (under 1000 ms), but is not recognized as an abnormality and is not reported. Finally, the remainder are decision errors in which abnormalities are fixated with a dwell time sufficient for decision making, but are rejected and hence not reported.

The key point to this classification of omission errors is that we look at the FNs as covert negative decisions. There is no overt response, but we assume that the prolonged dwell associated with them indicates an implicit decision not to respond. The classical signal detection framework has done well to advance our understanding of overt TP and FP responses, but has done little to aid our understanding of negative responses. Eye-position recording helps us understand errors of omission. At least 60% of FN decisions are associated with dwell times (766 to 1638 ms) significantly greater than dwell times associated with TN responses (460 to 892 ms); and much of the time, the FN dwells equal dwells associated with overt positive responses (495 to 2291 ms). Therefore, we can infer that there is a fairly significant amount of visual processing of the image at the FN location—sufficient to make a correct overt response as we show in this paper. Due to a failure in the recognition or decision process, however, the decision threshold for a positive response is not reached and the covert negative response is made.

Table 3 Percentages of FN and TN decisions having dwell times greater than 1000 ms for chest (tumors), mammography (microcalcifications and masses), bone trauma and bone fracture search. The rightmost column shows the ratio of FNs to TNs that would be highlighted during perceptual feedback at the 1000 ms threshold.

Image Type	FNs (%)	TNs (%)	FN/TN
Chest	59	18	3.3
Mammography	62	42	1.5
Bone trauma	44	37	1.2
Bone fracture	36	12	3.0

6 Perceptual Feedback

Perceptual feedback, a technique that was developed[11] to improve observer accuracy, capitalizes on the relationship between visual dwell and response outcomes. Perceptual feedback was designed to give radiologists a second chance to recognize missed lesions by utilizing the individual observer's own perceptual responses. Areas of prolonged dwell are highlighted with a circle immediately after an initial search of the image. The radiologist can review these areas and revise the initial decision. The hypothesis is that the second look at specific image locations will provide a chance for the observer to process the available image information knowing that they received prolonged dwell, indicating suspicious features. This unique perceptual feedback may be enough to shift the initial covert negative response to an overt (correct) positive response.

From the survival curves in Figure 3 it can be seen that in every case, if a 1000 ms threshold is used, a larger proportion of FN decisions than TN decisions have dwell greater than 1000 ms. Table 3 shows the percentages of FNs versus TNs that have dwell times greater than 1000 ms. For chest images, the 1000 ms threshold would provide a clear discrimination of FN and TN responses. The 1000 ms threshold would feedback about 59% of the FN image areas, but only about 18% of the TN areas for chest images. As Table 3 shows, the efficiency of the 1000 ms threshold varies for different types of images, so different thresholds may be required for successful perceptual feedback to work effectively using other images than chest. For example, with mammography, at 1000 ms the ratio of FNs to TNs is 1.5, but at 1200 ms it increases to 2.0.

There are a number of possible reasons why the numbers in Table 3 differ. Each experiment used a different type of radiographic image with different types of abnormalities to search for. Bone fracture and chest studies limited total search time to less than 30 s, while the bone trauma and mammography studies used unlimited search times. The unlimited search times increased the probability that the TN areas would be fixated more than once, driving up the TN dwell times compared to chest and fracture times. Driving up the TN dwell times increased the probability that these

Table 4 TP, FN, FP, and TN percentages for the feedback experiment—first look decisions, second look without feedback decisions, and second look with feedback decisions.

Condition	TP	FN	FP	TN
First look	35	65	8	92
No feedback	37	63	11	89
Feedback	54	46	10	90

areas had dwells greater than 1000 ms, so more TN areas would pass the perceptual feedback threshold, decreasing the FN/TN perceptual feedback ratio. This implies that the perceptual feedback algorithm may work most efficiently on only the first 30 s of search, ignoring search that occurs after the first 30 s.

For the perceptual feedback experiment[11] considered here, an algorithm was developed that analyzes the eye-position data and determines which image locations received cumulative clusters with dwell times exceeding 1000 ms. The observer was then given a second look at the image with 5 deg circles outlining the image locations receiving prolonged dwell. The observers could revise any decisions made during the first look at the image. The results were compared to a control condition in which observers were merely given a second look at the image without perceptual feedback circles provided. Six radiologist observers participated in the study, searching 40 chest images, 20 with one to three tumors and 20 without.

The decision data were analyzed using alternative free response operating characteristic (AFROC) techniques. The measure of performance in AFROC analysis is area under the AFROC curve or A1 (A1 ranges from 0 for chance performance to 1.0 for perfect performance). The A1 for the initial look at the image was 0.495 and 0.540, respectively, for the perceptual feedback and control conditions. After perceptual feedback or the second look without perceptual feedback, A1 was 0.618 and 0.504, respectively. For the control condition, the change in performance was not statistically significant. For the perceptual feedback condition, the nearly 20% improvement in performance was statistically significant using a t test for paired observations on the AFROC A1 results ($t=40.38$, df$=5$, $p<0.001$, where df indicates degrees of freedom). This difference indicates that perceptual feedback significantly improved recognition of tumors in chest images. The decision performance data are presented in Table 4. Note the nearly 20% increase in the TP rate for the feedback versus second look without feedback conditions.

Table 5 shows the dwell times associated with the various decision changes made with perceptual feedback. The change decisions (e.g., FN to TP) have dwells that fall intermediate between the positively maintained decisions (i.e., TP to TP, FP to FP) and the negatively maintained decisions (i.e., TN to TN, FN to FN). Speculation about the processing of negative decisions is inferred and are thus determined by default. The TN to TN combinations are based on the default decision not to report an abnormality-free image area as negative. These decisions would follow the negative path in Figure 2. The TN default decisions are typically fixated by one or more clusters of fixations that

Table 5 Dwell times in milliseconds associated with combinations of decisions made prior to perceptual feedback and decisions made with perceptual feedback. The fit categories refer to steps in the model presented in Figure 1. These results are based on data from Ref. 11.

Initial Decision	Decision after Perceptual Feedback	
	Positive	Negative
Positive	TP to TP 2382 ms	FP to TN 2199 ms
	FP to FP 2556 ms	TP to FN 2816 ms
	Mean=2469 ms	Mean=2247 ms
	Definite positive fit	Possible positive fit
Negative	TN to FP 1610 ms	TN to TN 787 ms
	FN to TP 1933 ms	FN to FN 1230 ms
	Mean=1710 ms	Mean=1008 ms
	Possible negative fit	Definite negative fit

required an average of 787 ms. The gaze duration of two out of three decisions that could be designated negative fit decisions in Figure 2 (TN to FP and FN to TP) changed to positive after reevaluation during the perceptual feedback view. Changing from an initially negative to a positive decision (true or false) added an average 763 ms of additional information processing to a default negative decision. Falsely maintaining a negative decision (FN to FN) added 443 ms of additional information processing to a default negative decision. This dwell variation presumably reflects a difference in information processing between fixating and recognizing or not recognizing an abnormality that is in the image. The increase in FN dwell may be required to disembed clutter obscuring the abnormality, but the point is that the region not reported contained a true abnormality and attracted prolonged visual dwell.

A series of follow-up experiments[17,18] to the perceptual feedback experiment demonstrated that perceptual feedback may be aiding abnormality recognition in two ways: by locating potential target areas and by enhancing the perception of targets. By comparing viewing of image locations with and without the circle highlight, it was found that when the circle was present, fixations tended to be less dispersed (0.89 versus 1.13 deg; $F(1,3)=14.43$, $p<0.05$; analysis of variance test) and they tended to fall directly on the abnormality more often (15 versus 8%) than when the circle was absent. The circle may be functioning as a fiducial marker for the visual-attention system, giving it a specifically bound region within which to focus or allocate its limited resources, producing a local perceptual enhancement effect.

7 Discussion

Why does perceptual feedback enhance performance to such a significant degree, when other methods of cueing in radiology [e.g., clinical history prompts,[19–21] checklists,[22] dual reading,[23] CAD (Refs. 24 to 29)] have reported equivocal results? One reason may be that perceptual feedback uses one decision maker to reevaluate cued locations based on the radiologist's own perceptual responses. Other methods, such as CAD and dual reading, use two independent decision makers and require a combination of deci-

sions from these two independent sources. For the radiologist, reviewing the CAD results or the opinion of another radiologist involves examining the image again and possibly paying attention to areas that were not considered in their own initial search of the image. This requires further information processing and in some cases a completely new decision to be made. With perceptual feedback, it is a reconsideration of a decision that the same radiologist had already made.

Cueing studies from the psychology literature[30–32] suggest another reason why perceptual feedback may work so well. Perceptual feedback provides a direct cue—a circle that is physically superimposed on the radiographic image. Cueing methods such as clinical history use indirect cues such as "check the third interspace on the chest," which do not physically change the image. The physical aspect of the perceptual feedback cue may be an important factor in why it works so effectively. A recent study by Cheal and Gregory[33] suggests that cueing not only facilitates target recognition, but it also reduces noise from other nontarget features in the display. The targets in this study were simple geometric shapes in a background of similar geometric shapes, but the same result was found in a study using radiographic images. Krupinski et al.[34] used chest images with tumor targets and demonstrated that cueing reduces significantly the effects of noise from nontarget features outside the region of the perceptual feedback circle. In fact, the feedback circle cue was so effective that if another tumor was located outside of the feedback circle, detection of the outside tumor was reduced significantly.

The studies presented here demonstrated that perceptual feedback can enhance significantly the detection and recognition of tumors in radiographic chest images. Based on the similarities in visual dwell data and the survival curves for lesions in mammography and fractures and trauma in bone images, it is quite likely that perceptual feedback will meet with the same success in these and other types of radiographic images. With advances in remote eye-position recording systems and techniques, perceptual feedback could find a place in the clinical environment.

Acknowledgments

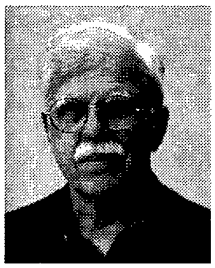
This work was supported in part by grants from the NCI, USPHS (CA-32870), the U.S. Army MRMC, the Department of Defense (BC-961120) and Toshiba Medical Systems, Tokyo, Japan.

References

1. H. L. Kundel, "Perception errors in chest radiography," *Semin. Respir. Med.* **10**, 203–210 (1989).
2. R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiology* **184**, 613–617 (1992).
3. R. T. Dahlen and H. T. Foley, "Medical malpractice claims in diagnostic radiology: update (letter)," *Radiology* **170**, 277 (1989).
4. R. S. Pillalamarri, B. D. Barnette, D. Birkmire, and R. Karsh, "Cluster: a program for the identification of eye-fixation-cluster characteristics," *Behav. Res. Meth. Instrum. Comput.* **25**, 9–15 (1993).
5. J. M. Henderson, K. K. McClure, S. Pierce, and G. Schrock, "Object identification without foveal vision: evidence from an artificial scotoma paradigm," *Percept. Psychophys.* **59**, 323–346 (1997).
6. G. R. Loftus and N. H. Mackworth, "Cognitive determinants of fixation location during picture viewing," *J. Exper. Psychol. Hum. Percept. Perform.* **4**, 565–572 (1978).
7. C. F. Nodine, H. L. Kundel, L. C. Toto, and E. A. Krupinski, "Recording and analyzing eye-position data using a microcomputer workstation," *Behav. Res. Meth. Instrum. Comput.* **24**, 475–485 (1992).
8. C. F. Nodine, H. L. Kundel, J. Polikoff, and L. Toto, "Using eye movements to study decision making of radiologists," in *Eye Movement Research: Physiological and Psychological Aspects*, G. Luer, U. Lass, J. Shallo-Hoffman, Eds., pp. 349–363, Hogrefe, Gottingen, Germany (1988).
9. H. L. Kundel, C. F. Nodine, D. Thickman, and L. Toto, "Searching for lung tumors: a comparison of human performance with random and systematic models," *Invest. Radiol.* **22**, 417–422 (1987).
10. C. F. Nodine and H. L. Kundel, "Computer-assisted perception aids pulmonary-nodule detection," in *Medical Imaging, Proc. SPIE* **2166**, 55–58 (1994).
11. H. L. Kundel, C. F. Nodine, and E. A. Krupinski, "Computer-displayed eye position as a visual aid to pulmonary tumor interpretation," *Invest. Radiol.* **25**, 890–896 (1990).
12. E. A. Krupinski, "Visual scanning patterns of radiologists searching mammograms," *Acad. Radiol.* **3**, 137–144 (1996).
13. E. A. Krupinski and P. J. Lund, "Differences in time to interpretation for evaluation of bone radiographs with monitor and film viewing," *Acad. Radiol.* **4**, 177–182 (1997).
14. C. Hu, H. L. Kundel, C. F. Nodine, E. A. Krupinski, and L. C. Toto, "Searching for bone fractures: a comparison with pulmonary tumor search," *Acad. Radiol.* **1**, 25–32 (1994).
15. R. C. Elandt-Johnson and N. L. Johnson, *Survival Models and Data Analysis*, John Wiley and Sons, New York (1980).
16. H. L. Kundel, C. F. Nodine, and D. P. Carmody, "Visual scanning, pattern recognition and decision-making in pulmonary tumor detection," *Invest. Radiol.* **13**, 175–181 (1978).
17. E. A. Krupinski, C. F. Nodine, and H. L. Kundel, "A perceptually based method for enhancing pulmonary tumor recognition," *Invest. Radiol.* **28**, 289–294 (1993).
18. E. A. Krupinski, C. F. Nodine, and H. L. Kundel, "Perceptual enhancement of tumor targets in chest x-ray images," *Percept. Psychophys.* **53**, 519–526 (1993).
19. K. S. Berbaum, E. A. Franken, K. L. Anderson, D. D. Dorfman, W. E. Erkonen, G. P. Farrar, J. J. Geraghty, T. J. Gleason, M. E. MacNaughton, M. E. Phillips, D. L. Renfrew, C. W. Walker, C. G. Whitten, and D. C. Young, "The influence of clinical history on visual search with single and multiple abnormalities," *Invest. Radiol.* **28**, 191–210 (1993).
20. U. O. Aideyan, K. Berbaum, and W. L. Smith, "Influence of prior radiologic information on the interpretation of radiographic examinations," *Acad. Radiol.* **2**, 205–208 (1995).
21. K. White, K. Berbaum, and W. L. Smith, "The role of previous radiographs and reports in the interpretation of current radiographs," *Invest. Radiol.* **29**, 263–265 (1994).
22. D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, "Enhanced interpretation of diagnostic images," *Invest. Radiol.* **23**, 240–252 (1988).
23. C. A. Beam, "Effect of human variability on independent double reading in screening mammography," *Acad. Radiol.* **3**, 891–897 (1996).
24. M. D. Mugglestone, R. Lomax, A. G. Gale, and A. R. M. Wilson, "The effect of prompting mammographic abnormalities on the human observer," in *Digital Mammography '96*, K. Doi, M. L. Giger, R. M. Nishikawa, R. A. Schmidt, Eds., pp. 87–92, Elsevier, New York (1996).
25. M. L. Giger, "Computer-aided diagnosis," in *A Categorical Course in Physics. Technical Aspects of Breast Imaging*, A. G. Haus and M. J. Yaffe, Eds., pp. 272–298, RSNA Publications, Oak Brook, IL (1993).
26. W. Zhang, K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, "An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms," *Med. Phys.* **23**, 595–601 (1996).
27. N. F. Vittitoe, J. A. Baker, and C. E. Floyd, "Fractal texture analysis in computer-aided diagnosis of solitary pulmonary tumors," *Acad. Radiol.* **4**, 96–101 (1997).
28. H. Yoshida, K. Doi, R. M. Nishikawa, M. L. Giger, and R. A. Schmidt, "An improved computer-assisted diagnostic scheme using wavelet transform for detecting clustered microcalcifications in digital mammograms," *Acad. Radiol.* **3**, 621–627 (1996).
29. E. A. Krupinski and R. M. Nishikawa, "Comparison of eye position versus computer identified microcalcification clusters on mammograms," *Med. Phys.* **24**, 17–23 (1997).
30. B. J. A. Krose and B. Julesz, "The control and speed of shifts of attention," *Vis. Res.* **29**, 1607–1619 (1989).
31. C. W. Eriksen and Y. Yeh, "Allocation of attention in the visual field," *J. Exper. Psychol. Hum. Percept. Perform.* **11**, 583–597 (1986).
32. G. Chastain, M. Cheal, and D. R. Lyon, "Attention and nontarget effects in the location-cueing paradigm," *Percept. Psychophys.* **58**, 300–309 (1996).
33. M. L. Cheal and M. Gregory, "Evidence of limited capacity and noise-reduction with single-element displays in the location-cueing paradigm," *J. Exper. Psychol. Hum. Percept. Perform.* **23**, 51–71 (1997).

34. E. A. Krupinski, C. F. Nodine, and H. L. Kundel, "Perceptual enhancement of tumor targets in chest x-ray images," *Percept. Psychophys.* **53**, 519–526 (1993).

Elizabeth A. Krupinski received her undergraduate degree in psychology from Cornell University and her MA and PhD degrees in experimental psychology from Montclair State College and Temple University. She was a research specialist for 5 years with the University of Pennsylvania Department of Radiology and is currently a research associate professor with the Departments of Radiology and Psychology at the University of Arizona, where she has been for 5 years. Her main interest is understanding the perceptual and decision-making strategies of radiologists searching images for lesions, and using this information to understand and reduce errors in radiology. In addition to general issues of observer performance, Dr. Krupinski is interested in performance and ergonomic issues associated with reading radiologic images from workstations.



Calvin F. Nodine received his BA and MA degrees in psychology from Bucknell University in 1954 and 1959, respectively, and his PhD in experimental psychology from the University of Massachusetts in 1962. Dr. Nodine is currently a research professor of radiologic science with the University of Pennsylvania. He is currently working on computer-assisted perception of medical images.

Harold L. Kundel received his AB and MD from Columbia University and his residency training in radiology from Temple University. During a fellowship in the Radiology-Physiology Laboratory at Temple University, he became interested in eye-position recording as a method for studying the source of perceptual error in radiology. This interest broadened into studies of the use of observer performance methodology, including statistical decision theory, for the evaluation of emerging digital imaging technology. He is currently the Matthew Wilson Professor of Research Radiology at the University of Pennsylvania where he is working on modeling and evaluating picture archiving and communication systems as well as visual search.